Multimodal Task-Driven Dictionary Learning for Image Classification

Soheil Bahrampour, *Student Member, IEEE,* Nasser M. Nasrabadi, *Fellow, IEEE,* Asok Ray, *Fellow, IEEE,* and W. Kenneth Jenkins, *Life Fellow, IEEE*

Abstract—**Dictionary learning algorithms have been successfully used for both reconstructive and discriminative tasks, where an input signal is represented with a sparse linear combination of dictionary atoms. While these methods are mostly developed for single-modality scenarios, recent studies have demonstrated the advantages of feature-level fusion based on the joint sparse representation of the multimodal inputs. In this paper, we propose a *multimodal* task-driven dictionary learning algorithm under the joint sparsity constraint (prior) to enforce collaborations among multiple homogeneous/heterogeneous sources of information. In this task-driven formulation, the multimodal dictionaries are learned simultaneously with their corresponding classifiers. The resulting multimodal dictionaries can generate discriminative latent features (sparse codes) from the data that are optimized for a given task such as binary or multiclass classification. Moreover, we present an extension of the proposed formulation using a mixed joint and independent sparsity prior which facilitates more flexible fusion of the modalities at feature level. The efficacy of the proposed algorithms for multimodal classification is illustrated on four different applications – multimodal face recognition, multi-view face recognition, multi-view action recognition, and multimodal biometric recognition. It is also shown that, compared to the counterpart reconstructive-based dictionary learning algorithms, the task-driven formulations are more computationally efficient in the sense that they can be equipped with more compact dictionaries and still achieve superior performance.**

Index Terms—**Dictionary learning, Multimodal classification, Sparse representation, Feature fusion**

I. INTRODUCTION

It is well established that information fusion using multiple sensors can generally result in an improved recognition performance [1]. It provides a framework to combine local information from different perspectives which is more tolerant to the errors of individual sources [2], [3]. Fusion methods for classification are generally categorized into feature fusion [4] and classifier fusion [5] algorithms. Feature fusion methods aggregate extracted features from different sources into a single feature set which is then used for classification. On the other hand, classifier fusions algorithms combine decisions from individual classifiers, each of which is trained using separate sources. While classifier fusion is a well-studied topic, fewer studies have been done

S. Bahrampour, A. Ray, and W. K. Jenkins are with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802, USA; {soheil, axr2, wkj1}@psu.edu.

N. M. Nasrabadi is with the Army Research Laboratory, Adelphi, MD 20783; nasser.m.nasrabadi.civ@mail.mil.

for feature fusion, mainly due to the incompatibility of the feature sets [6]. A naive way of feature fusion is to stack the features into a longer one [7]. However this approach usually suffers from the curse of dimensionality due to the limited number of training samples [4]. Even in scenarios with abundant training samples, concatenation of feature vectors does not take into account the relationship among the different sources and it may contain noisy or redundant data, which degrade the performance of the classifier [6]. However, if these limitations are mitigated, feature fusion can potentially result in improved classification performance [8], [9].

Sparse representation classification has recently attracted the interest of many researchers in which the input signal is approximated with a linear combination of *a few* dictionary atoms [10] and has been successfully applied to several problems such as robust face recognition [10], visual tracking [11], and transient acoustic signal classification [12]. In this approach, a structured dictionary is usually constructed by stacking all the training samples from the different classes. The method has also been expanded for efficient feature-level fusion which is usually referred to as multi-task learning [13], [14], [15], [16]. Among different proposed sparsity constraints (priors), joint sparse representation has shown significant performance improvement in several multi-task learning applications such as target classification, biometric recognitions, and multiview face recognition [12], [14], [17], [18]. The underlying assumption is that the multimodal test input can be simultaneously represented by a few dictionary atoms, or training samples, from a multimodal dictionary, that represents all the modalities and, therefore, the resulting sparse coefficients should have the same sparsity pattern. However, the dictionary constructed by the collection of the training samples suffer from two limitations. First, as the number of training samples increases, the resulting optimization problem becomes more computationally demanding. Second, the dictionary that is constructed this way is not optimal neither for the reconstructive tasks [19] nor the discriminative tasks [20].

Recently it has been shown that *learning the dictionary* can overcome the above limitations and significantly improve the performance in several applications including image restoration [21], face recognition [22] and object recognition [23], [24]. The learned dictionaries are usually more compact and have fewer dictionary atoms than the number of training samples [25], [26]. Dictionary

learning algorithms can generally be categorized into two groups: unsupervised and supervised. Unsupervised dictionary learning algorithms such as the method of optimal direction [27] and K-SVD [25] are aimed at finding a dictionary that yields minimum errors when adapted to reconstruction tasks such as signal denoising [28] and image inpainting [19]. Although, the unsupervised dictionary learning has also been used for classification [22], it has been shown that better performance can be achieved by learning the dictionaries that are adapted to an specific task rather than just the data set [29], [30]. These methods are called supervised, or *task-driven*, dictionary learning algorithms. For the classification task, for example, it is more meaningful to utilize the labeled data to minimize the misclassification error rather than the reconstruction error [31]. Adding a discriminative term to the reconstruction error and minimizing a trade-off between them has been proposed in several formulations [20], [24], [32], [33]. The incoherent dictionary learning algorithm proposed in [34] is another supervised formulation which trains class-specific dictionaries to minimize atom sharing between different classes and uses sparse representation for classification. In [35], a Fisher criterion is proposed to learn structured dictionaries such that the sparse coefficients have small within-class and large between-class scatters. While unsupervised dictionary learning can be reformulated as a large scale matrix factorization problem and solved efficiently [19], supervised dictionary learning is usually more difficult to optimize. More recently, it has been shown that better optimization tool can be used to tackle the supervised dictionary learning [30], [36]. This is achieved by formulating it as a bilevel optimization problem [37], [38]. In particular, a stochastic gradient descent algorithm has been proposed in [29] which efficiently solves the dictionary learning problem in a unified framework for different tasks, such as classification, nonlinear image mapping, and compressive sensing.

The majority of the existing dictionary learning algorithms, including the task-driven dictionary learning [29], are only applicable to single source of data. In [39], a set of view-specific dictionaries and a common dictionary are learned for the application of multi-view action recognition. The view-specific dictionaries are trained to exploit view-level correspondence while the common dictionary is trained to capture common patterns shared among the different views. The proposed formulation belongs to the class of dictionary learning algorithms that leverages the labeled samples to learn class-specific atoms while minimizing the reconstruction error. Moreover, it cannot be used for fusion of the heterogeneous modalities. In [40], a generative multimodal dictionary learning algorithm is proposed to extract typical templates of multimodal features. The templates represent synchronous transient structures between modalities which can be used for localization applications. More recently, a multimodal dictionary learning algorithm with joint sparsity prior is proposed in [41] for multimodal retrieval where the task is to find relevant samples from other modalities for a given unimodal query. However,
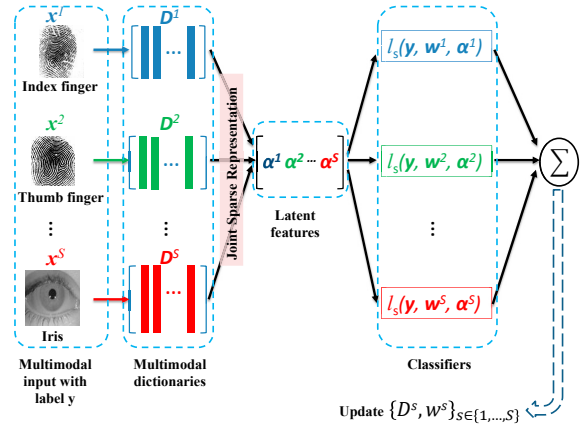

Fig. 1: Multimodal task-driven dictionary learning scheme.

the proposed formulation cannot be readily applied for information fusion in which the task is to find label of a given multimodal query. Moreover, the joint sparsity prior is used in [41] to couple similarly labeled samples within each modality and is not utilized to extract cross-modality information which is essential for information fusion [12]. Furthermore, the dictionaries in [41] are learned to be generative by minimizing the reconstruction error of data across modalities and, therefore, are not necessary optimal for discriminative tasks [31].

This paper focuses on learning discriminative multimodal dictionaries. The major contributions of the paper are as follows:

- *Formulation of the multimodal dictionary learning algorithms*: A multimodal task-driven dictionary learning algorithm is proposed for classification using homogeneous or heterogeneous sources of information. Information from different modalities are fused both at the feature level, by using the joint sparse representation, and at the decision level, by combining the scores of the modal-based classifiers. The proposed formulation simultaneously trains the multimodal dictionaries and classifiers under the joint sparsity prior in order to enforce collaborations among the modalities and obtain the latent sparse codes as the optimized features for different tasks such as binary and multiclass classification. Fig. 1 presents an overview of the proposed framework. An unsupervised multimodal dictionary learning algorithm is also presented as a by-product of the supervised version.

- *Differentiability of the bi-level optimization problem*: The main difficulty in proposing such a formulation is that the solution of the corresponding joint sparse coding problem is not differentiable with respect to the dictionaries. While the joint sparse coding has a non-smooth cost function, it is shown here that it is locally differentiable and the resulting bi-level optimization for task-driven multimodal dictionary learning is smooth and can be solved using a stochastic

gradient descent algorithm. [1]

- *Flexible feature-level fusion*: An extension of the proposed framework is presented which facilitates more flexible fusion of the modalities at the feature level by allowing the modalities to have different sparsity patterns. This extension provides a framework to tune the trade-off between independent sparse representation and joint sparse representation among the modalities.

- *Improved performance for multimodal classification*: The proposed methods achieve the state-of-the-art performance in a range of different multi-modal classification tasks. In particular, we have provided extensive performance comparison between the proposed algorithms and some of the competing methods from literature for four different tasks of multimodal face recognition, multi-view face recognition, multimodal biometric recognition, and multi-view action recognition. The experimental results on these datasets have demonstrated the usefulness of the proposed formulation, showing that the proposed algorithm can be readily applied to several different application domains.

- *Improved efficiency for sparse-representation based classification*: It is shown here that, compared to the counterpart sparse representation classification algorithms, the proposed algorithms are more computationally efficient in the sense that they can be equipped with more compact dictionaries and still achieve superior performance.

A. Paper organization

The rest of the paper is organized as follows. In Section II, unsupervised and supervised dictionary learning algorithms for single source of information are reviewed. Joint sparse representation for multimodal classification is also reviewed in this section. Section III proposes the task-driven multimodal dictionary learning algorithms. Comparative studies on several benchmarks and concluding results are presented in Section IV and Section V, respectively.

B. Notation

Vectors are denoted by bold lower case letters and matrices by bold upper case letters. For a given vector $\boldsymbol{x}$, x_i is its ith element. For a given finite set of indices γ, $\boldsymbol{x}_\gamma$ is the vector formed with those elements of $\boldsymbol{x}$ indexed in γ. Symbol $\rightarrow$ is used to distinguish the row vectors from column vectors, i.e. for a given matrix $\boldsymbol{X}$, the ith row and j^{th} column of matrix are represented as $\boldsymbol{x}_{i\rightarrow}$ and $\boldsymbol{x}_j$, respectively. For a given finite set of indices γ, $\boldsymbol{X}_\gamma$ is the matrix formed with those columns of $\boldsymbol{X}$ indexed in γ and $\boldsymbol{X}_{\gamma\rightarrow}$ is the matrix formed with those rows of $\boldsymbol{X}$ indexed in γ. Similarly, for given finite sets of indices γ and ψ, $\boldsymbol{X}_{\gamma\rightarrow,\psi}$ is the matrix formed with those rows and columns of $\boldsymbol{X}$ indexed in γ and ψ, respectively. x_{ij} is the element of $\boldsymbol{X}$ at row i and column j. The l_q norm, $q \geq 1$, of a vector $\boldsymbol{x} \in \mathbb{R}^m$ is defined as $\|\boldsymbol{x}\|_{\ell_q} = \left(\sum_{j=1}^m |x_j|^q\right)^{1/q}$. The Frobenius norm and ℓ_{1q} norm, $q \geq 1$, of matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ is defined as $\|\boldsymbol{X}\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n x_{ij}^2\right)^{1/2}$ and $\|\boldsymbol{X}\|_{\ell_{1q}} = \sum_{i=1}^m \|\boldsymbol{x}_{i\rightarrow}\|_{\ell_q}$, respectively. The collection $\{\boldsymbol{x}^i | i \in \gamma\}$ is shortly denoted as $\{\boldsymbol{x}^i\}$.

II. BACKGROUND

A. Dictionary learning

Dictionary learning has been widely used in various tasks such as reconstruction, classification, and compressive sensing [29], [33], [42], [43]. In contrast to principal component analysis (PCA) and its variants, dictionary learning algorithms generally do not impose orthogonality condition and are more flexible allowing to be well-tuned to the training data. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{n\times N}$ be the collection of N (normalized) training samples that are assumed to be statistically independent. Dictionary $\boldsymbol{D} \in \mathbb{R}^{n\times d}$ can then be obtained as the minimizer of the following empirical cost [22]:

$$g_N(\boldsymbol{D}) \triangleq \frac{1}{N}\sum_{i=1}^N l_u(\boldsymbol{x}_i, \boldsymbol{D}) \qquad (1)$$

over the regularizing convex set $\mathcal{D} \triangleq \{\boldsymbol{D} \in \mathbb{R}^{n\times d} | \|\boldsymbol{d}_k\|_{\ell_2} \leq 1, \forall k = 1, \ldots, d\}$, where $\boldsymbol{d}_k$ is the kth column, or atom, in the dictionary and the unsupervised loss l_u is defined as

$$l_u(\boldsymbol{x}, \boldsymbol{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{D}\boldsymbol{\alpha}\|_{\ell_2}^2 + \lambda_1\|\boldsymbol{\alpha}\|_{\ell_1} + \lambda_2\|\boldsymbol{\alpha}\|_{\ell_2}^2, \quad (2)$$

which is the optimal value of the sparse coding problem with λ_1 and λ_2 being the regularizing parameters. While λ_2 is usually set to zero to exploit sparsity, using $\lambda_2 > 0$ makes the optimization problem in Eq. (2) strongly convex resulting in a differentiable cost function [29]. The index u of l_u is used to emphasize that the above dictionary learning formulation is an unsupervised method. It is well-known that one is often interested in minimizing an expected risk, rather than the perfect minimization of the empirical cost [44]. An efficient online algorithm is proposed in [19] to find the dictionary $\boldsymbol{D}$ as the minimizer of the following stochastic cost over the convex set $\mathcal{D}$:

$$g(\boldsymbol{D}) \triangleq \mathrm{E}_{\boldsymbol{x}}[l_u(\boldsymbol{x}, \boldsymbol{D})], \qquad (3)$$

where it is assumed that the data $\boldsymbol{x}$ is drawn from a finite probability distribution $p(\boldsymbol{x})$ which is usually unknown and $\mathrm{E}_{\boldsymbol{x}}[.]$ is the expectation operator with respect to the distribution $p(\boldsymbol{x})$.

The trained dictionary can then be used to (sparsely) reconstruct the input. The reconstruction error has been shown to be a robust measure for classification tasks [10], [45]. Another use of a given trained dictionary is for feature extraction where the sparse code $\boldsymbol{\alpha}^\star(\boldsymbol{x}, \boldsymbol{D})$, obtained as a solution of (2), is used as a feature vector representing the input signal $\boldsymbol{x}$ in the classical expected risk optimization for training a classifier [29]:

$$\min_{\boldsymbol{w}\in\mathcal{W}} \mathrm{E}_{y,\boldsymbol{x}}[l(y, \boldsymbol{w}, \boldsymbol{\alpha}^\star(\boldsymbol{x}, \boldsymbol{D}))] + \frac{\nu}{2}\|\boldsymbol{w}\|_{\ell_2}^2, \qquad (4)$$

---

[1]The source code of the proposed algorithm is released here: https://github.com/soheilb/multimodal_dictionary_learning

where y is the ground truth class label associated with the input x, w is model (classifier) parameters, ν is a regularizing parameter, and l is a convex loss function that measures how well one can predict y given the feature vector $\alpha^\star$ and classifier parameters w. The expectation $\mathrm{E}_{y,x}$ is taken with respect to the probability distribution $p(y, x)$ of the labeled data. Note that in Eq. 4, the dictionary D is fixed and independent of the given task and class label y. In task-driven dictionary learning, on the other hand, a *supervised* formulation is used which finds the optimal dictionary and classifier parameters jointly by solving the following optimization problem [29]:

$$\min_{D \in \mathcal{D}, w \in \mathcal{W}} \mathrm{E}_{y,x} \left[l_{su} \left(y, w, \alpha^\star(x, D) \right) \right] + \frac{\nu}{2} \| w \|_{\ell_2}^2. \quad (5)$$

The index su of convex loss function l_{su} is used to emphasize that the above dictionary learning formulation is supervised. The learned task-driven dictionary has been shown to result in a superior performance compared to the unsupervised setting [29]. In this setting, the sparse codes are indeed the optimized latent features for the classifier.

B. Multimodal joint sparse representation

Joint sparse representation provides an efficient tool for feature-level fusion of sources of information [12], [14], [46]. Let $\mathcal{S} \triangleq \{1, \ldots, S\}$ be a finite set of available modalities and let $x^s \in \mathbb{R}^{n^s}, s \in \mathcal{S}$, be the feature vector for the sth modality. Also let $D^s \in \mathbb{R}^{n^s \times d}$ be the corresponding dictionary for the sth modality. For now, it is assumed that the multimodal dictionaries are constructed by collections of the training samples from different modalities, i.e. j^{th} atom of dictionary D^s is the jth training sample from the sth modality. Given a multimodal input $\{x^s | s \in \mathcal{S}\}$, shortly denoted as $\{x^s\}$, an optimal sparse matrix $A^\star \in \mathbb{R}^{d \times S}$ is obtained by solving the following ℓ_{12}-regularized reconstruction problem:

$$\underset{A=[\alpha^1 \ldots \alpha^S]}{\operatorname{argmin}} \frac{1}{2} \sum_{s=1}^{S} \| x^s - D^s \alpha^s \|_{\ell_2}^2 + \lambda \| A \|_{\ell_{12}}, \quad (6)$$

where λ is a regularization parameter. Here α^s is the sth-column of A which corresponds to the sparse representation for the sth modality. Different algorithms have been proposed to solve the above optimization problem [47], [48]. We use the efficient alternating direction method of multipliers (ADMM) [49] to find $A^\star$. The ℓ_{12} prior encourages row sparsity in $A^\star$, i.e. it encourages collaboration among all the modalities by enforcing the same dictionary atoms from different modalities that present the same event, to be used for reconstructing the inputs $\{x^s\}$. An ℓ_{11} term can also be added to the above cost function to extend it to a more general framework where sparsity can also be sought within the rows, as will be discussed in Section III-D. It has been shown that joint sparse representation can result in a superior performance in fusing multimodal sources of information compared to other information fusion techniques [45]. We are interested in learning multimodal dictionaries under the joint sparsity prior. This

has several advantages over a fixed dictionary consisting of training data. Most importantly, it can potentially remove the redundant and noisy information by representing the training data in a more compact form. Also using the supervised formulation, one expects to find dictionaries that are well-adapted to the discriminative tasks.

III. MULTIMODAL DICTIONARY LEARNING

In this section, online algorithms for unsupervised and supervised multimodal dictionary learning are proposed.

A. Multimodal unsupervised dictionary learning

Unsupervised multimodal dictionary learning is derived by extending the optimization problem characterized in Eq. (3) and using the joint sparse representation of (6) to enforce collaborations among modalities. Let the minimum cost $l'_u(\{x^s, D^s\})$ of the joint sparse coding be defined as

$$\min_A \frac{1}{2} \sum_{s=1}^{S} \| x^s - D^s \alpha^s \|_{\ell_2}^2 + \lambda_1 \| A \|_{\ell_{12}} + \frac{\lambda_2}{2} \| A \|_F^2, \quad (7)$$

where λ_1 and λ_2 are the regularizing parameters. The additional Frobenius norm $\|.\|_F$ compared to Eq. (6) guarantees a unique solution for the joint sparse optimization problem. In the special case when $S = 1$, optimization (7) reduces to the well-studied elastic-net optimization [50]. By natural extension of the optimization problem (3), the unsupervised multimodal dictionaries are obtained by:

$$D^{s\star} = \underset{D^s \in \mathcal{D}^s}{\operatorname{argmin}} \mathrm{E}_{x^s} \left[l'_u(\{x^s, D^s\}) \right], \forall s \in \mathcal{S}, \quad (8)$$

where the convex set $\mathcal{D}^s$ is defined as

$$\mathcal{D}^s \triangleq \{ D \in \mathbb{R}^{n^s \times d} | \| d_k \|_{\ell_2} \leq 1, \forall k = 1, \ldots, d \}. \quad (9)$$

It is assumed that data x^s is drawn from a finite (unknown) probability distribution $p(x^s)$. The above optimization problem can be solved using the classical projected stochastic gradient algorithm [51] which consists of a sequence of updates as follows:

$$D^s \leftarrow \Pi_{\mathcal{D}^s} \left[D^s - \rho_t \nabla_{D^s} l'_u(\{x^s_t, D^s\}) \right], \quad (10)$$

where ρ_t is the gradient step at time t and $\Pi_{\mathcal{D}}$ is the orthogonal projector onto set $\mathcal{D}$. The algorithm converges to a stationary point for a decreasing sequence of ρ_t [51], [52]. A typical choice of ρ_t is shown in the next section. This problem can also be solved using online matrix factorization algorithm [26]. It should be noted that the while the stochastic gradient descent does converge, it is not guaranteed to converge to a global minimum due to the non-convexity of the optimization problem [26], [44]. However, such stationary point is empirically found to be sufficiently good for practical applications [21], [28].

B. Multimodal task-driven dictionary learning

As discussed in Section II, the unsupervised setting does not take into account the label of the training data, and the dictionaries are obtained by minimizing the reconstruction error. However, for classification tasks, the minimum reconstruction error does not necessarily result in discriminative dictionaries. In this section, a multimodal task-driven dictionary learning algorithm is proposed that enforces collaboration among the modalities both at the feature level using joint sparse representation and the decision level using a sum of the decision scores. We propose to learn the dictionaries $\boldsymbol{D}^{s\star}, \forall s \in \mathcal{S}$, and the classifier parameters $\boldsymbol{w}^{s\star}, \forall s \in \mathcal{S}$, shortly denoted as the set $\{\boldsymbol{D}^{s\star}, \boldsymbol{w}^{s\star}\}$, jointly as the solution of the following optimization problem:

$$\min_{\{\boldsymbol{D}^s \in \mathcal{D}^s, \boldsymbol{w}^s \in \mathcal{W}^s\}} f\left(\{\boldsymbol{D}^s, \boldsymbol{w}^s\}\right) + \frac{\nu}{2} \sum_{s=1}^{S} \|\boldsymbol{w}^s\|_{\ell_2}^2, \quad (11)$$

where f is defined as the expected cumulative cost:

$$f\left(\{\boldsymbol{D}^s, \boldsymbol{w}^s\}\right) = \mathrm{E} \sum_{s=1}^{S} l_{su}(y, \boldsymbol{w}^s, \boldsymbol{\alpha}^{s\star}), \quad (12)$$

where $\boldsymbol{\alpha}^{s\star}$ is the sth column of the minimizer $\boldsymbol{A}^\star(\{\boldsymbol{x}^s, \boldsymbol{D}^s\})$ of the optimization problem (7) and $l_{su}(y, \boldsymbol{w}, \boldsymbol{\alpha})$ is a convex loss function that measures how well the classifier parametrized by $\boldsymbol{w}$ can predict y by observing $\boldsymbol{\alpha}$. The expectation is taken with respect to the joint probability distribution of the multimodal inputs $\{\boldsymbol{x}^s\}$ and label y. Note that $\boldsymbol{\alpha}^{s\star}$ acts as a hidden/latent feature vector, corresponding to the input $\boldsymbol{x}^s$, which is generated by the learned discriminative dictionary $\boldsymbol{D}^{s\star}$. In general, l_{su} can be chosen as any convex function such that $l_{su}(y, ., .)$ is twice continuously differentiable for all possible values of y. A few examples are given below for binary and multiclass classification tasks.

1) Binary classification: In a binary classification task where the label y belongs to the set $\{-1, 1\}$, l_{su} can be naturally chosen as the logistic regression loss

$$l_{su}(y, \boldsymbol{w}, \boldsymbol{\alpha}^\star) = \log(1 + e^{-y\boldsymbol{w}^T\boldsymbol{\alpha}^\star}), \quad (13)$$

where $\boldsymbol{w} \in \mathbb{R}^d$ is the classifier parameters. Once the optimal $\{\boldsymbol{D}^s, \boldsymbol{w}^s\}$ are obtained, a new multimodal sample $\{\boldsymbol{x}^s\}$ is classified according to sign of $\sum_{s=1}^{S} \boldsymbol{w}^{sT}\boldsymbol{\alpha}^\star$ due to the uniform monotonicity of $\sum_{s=1}^{S} l_{su}$. For simplicity, the intercept term for the linear model is omitted here, but it can be easily added. One can also use a bilinear model where, instead of a set of vectors $\{\boldsymbol{w}^s\}$, a set of matrices $\{\boldsymbol{W}^s\}$ are learned and a new multimodal sample is classified according to the sign of $\sum_{s=1}^{S} \boldsymbol{x}^{sT}\boldsymbol{W}^s\boldsymbol{\alpha}^\star$. Accordingly, the ℓ_2-norm regularization of Eq. (11) needs to be replaced with the matrix Frobenius norm. The bilinear model is richer than the linear model and can sometimes result in better classification performance but needs more careful training to avoid over-fitting.

2) Multiclass classification: Multiclass classification can be formulated using a collections of (independently learned) binary classifiers in a one-vs-one or one-vs-all setting. Multiclass classification can also be handled in an all-vs-all setting using the softmax regression loss function. In this scheme, the label y belongs to the set $\{1, \ldots, K\}$ and the softmax regression loss is defined as

$$l_{su}(y, \boldsymbol{W}, \boldsymbol{\alpha}^\star) = -\sum_{k=1}^{K} 1_{\{y=k\}} \log\left(\frac{e^{\boldsymbol{w}_k^T\boldsymbol{\alpha}^\star}}{\sum_{l=1}^{K} e^{\boldsymbol{w}_l^T\boldsymbol{\alpha}^\star}}\right), \quad (14)$$

where $\boldsymbol{W} = [\boldsymbol{w}_1 \ldots \boldsymbol{w}_K] \in \mathbb{R}^{d \times K}$, and $1_{\{.\}}$ is the indicator function. Once the optimal $\{\boldsymbol{D}^s, \boldsymbol{W}^s\}$ are obtained, a new multimodal sample $\{\boldsymbol{x}^s\}$ is classified as

$$\mathrm{argmax}_{k \in \{1, \ldots, K\}} \sum_{s=1}^{S} \left(\frac{e^{\boldsymbol{w}_k^{sT}\boldsymbol{\alpha}^{s\star}}}{\sum_{l=1}^{K} e^{\boldsymbol{w}_l^{sT}\boldsymbol{\alpha}^{s\star}}}\right). \quad (15)$$

In yet another all-vs-all setting, the multiclass classification task can be turned into a regression task in which the scaler label y is changed to a binary vector $\boldsymbol{y} \in \mathbb{R}^K$, where the kth coordinate corresponding to the label of $\{\boldsymbol{x}^s\}$ is set to one and the rest of the coordinates are set to zero. In this setting, l_{su} is defined as

$$l_{su}(\boldsymbol{y}, \boldsymbol{W}, \boldsymbol{\alpha}^\star) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\alpha}^\star\|_{\ell_2}^2, \quad (16)$$

where $\boldsymbol{W} \in \mathbb{R}^{K \times d}$. Having obtained the optimal $\{\boldsymbol{D}^s, \boldsymbol{W}^s\}$, the test sample $\{\boldsymbol{x}^s\}$ is then classified as

$$\mathrm{argmin}_{k \in \{1, \ldots, K\}} \sum_{s=1}^{S} \|\boldsymbol{q}^k - \boldsymbol{W}^s\boldsymbol{\alpha}^{s\star}\|_{\ell_2}^2, \quad (17)$$

where $\boldsymbol{q}^k$ is a binary vector in which its kth coordinate is one and its remaining coordinates are zero.

In choosing between the one-vs-all setting, in which independent multimodal dictionaries are trained for each class, and the multiclass formulation, in which multimodal dictionaries are shared between classes, a few points should be considered. In the one-vs-all setting, the total number of dictionary atoms is equal to dSK in the K-class classification while in the multiclass setting the number is equal to dS. It should be noted that in the multiclass setting a larger dictionary is generally required to achieve the same level of performance to capture the variations among all classes. However, it is generally observed that the size of the dictionaries in multiclass setting is not required to grow linearly as the number of classes increases due to atom sharing among the different classes. Another point to consider is that the class-specific dictionaries of the one-vs-all approach are independent and can be obtained in parallel. In this paper, the multiclass formulation is used to allow feature sharing among the classes.

C. Optimization

The main challenge in optimizing (11) is the non-differentiability of $\boldsymbol{A}^\star(\{\boldsymbol{x}^s, \boldsymbol{D}^s\})$. However, it can be shown that although the sparse coefficients $\boldsymbol{A}^\star$ are obtained by solving a non-differentiable optimization problem, the

function $f(\{\boldsymbol{D}^s, \boldsymbol{w}^s\})$, defined in Eq. (12), is differentiable on $\mathcal{D}^1 \times \cdots \mathcal{D}^S \times \mathcal{W}^1 \times \cdots \mathcal{W}^S$, and therefore its gradients are computable. To find the gradient of f with respect to $\boldsymbol{D}^s$, one can find the optimality condition of the optimization (7) or use the fixed point differentiation [36], [38] and show that $\boldsymbol{A}^\star$ is differentiable over its non-zero rows. Without loss of generality, we assume that label y admits a finite set of values such as those defined in Eqs. (13) and (14). The same algorithm can be derived for the scenario when $\boldsymbol{y}$ belongs to a compact subset of a finite-dimensional real vector space as in Eq. (16). A couple of mild assumptions are required to prove the differentiability of f which are direct generalizations of those required for the single modal scenario [29] and are listed below:

Assumption (A). The multimodal data $(y, \{\boldsymbol{x}^s\})$ admit a probability density p with compact support.

Assumption (B). For all possible values of y, $p(y, .)$ is continuous and $l_{su}(y, .)$ is twice continuously differentiable.

The first assumption is reasonable when dealing with the signal/image processing applications where the acquired values obtained by the sensors are bounded. Also all the given examples for l_{su} in the previous section satisfy the second assumption. Before stating the main proposition of this paper below, the term *active set* is defined.

Definition 3.1 (Active set): The active set Λ of the solution $\boldsymbol{A}^\star$ of the joint sparse coding problem (7) is defined to be

$$\Lambda = \{j \in \{1, \ldots, d\} : \|\boldsymbol{a}^\star_{j\rightarrow}\|_{\ell_2} \neq 0\}, \tag{18}$$

where $\boldsymbol{a}^\star_{j\rightarrow}$ is the jth row of $\boldsymbol{A}^\star$.

Proposition 3.1 (Differentiability and gradients of f): Let $\lambda_2 > 0$ and the assumptions (A) and (B) hold. Let $\Upsilon = \cup_{j\in\Lambda}\Upsilon_j$ where $\Upsilon_j = \{j, j + d, \ldots, j + (S - 1)d\}$. Let the matrix $\hat{\boldsymbol{D}} \in \mathbb{R}^{n\times|\Upsilon|}$ be defined as

$$\hat{\boldsymbol{D}} = \left[\hat{\boldsymbol{D}}_1 \ldots \hat{\boldsymbol{D}}_{|\Lambda|} \right], \tag{19}$$

where $\hat{\boldsymbol{D}}_j = \mathrm{blkdiag}(\boldsymbol{d}^1_j, \ldots, \boldsymbol{d}^S_j) \in \mathbb{R}^{n\times S}, \forall j \in \Lambda$, is the collection of the jth active atoms of the multimodal dictionaries, $\boldsymbol{d}^s_j$ is the jth active atom of $\boldsymbol{D}^s$, blkdiag is the block diagonalization operator, and $n = \sum_{s\in\mathcal{S}} n^s$. Also let matrix $\boldsymbol{\Delta} \in \mathbb{R}^{|\Upsilon|\times|\Upsilon|}$ be defined as

$$\boldsymbol{\Delta} = \mathrm{blkdiag}(\boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_{|\Lambda|}), \tag{20}$$

where $\boldsymbol{\Delta}_j = \frac{1}{\|\boldsymbol{a}^\star_{j\rightarrow}\|_{\ell_2}}\boldsymbol{I} - \frac{1}{\|\boldsymbol{a}^\star_{j\rightarrow}\|_{\ell_2}^3}\boldsymbol{a}^{\star\,T}_{j\rightarrow}\boldsymbol{a}^\star_{j\rightarrow} \in \mathbb{R}^{S\times S}, \forall j \in \Lambda$, and $\boldsymbol{I}$ is the identity matrix. Then, the function f defined in Eq. (12) is differentiable and $\forall s \in \mathcal{S}$,

$$\nabla_{\boldsymbol{w}^s} f = \mathrm{E}\left[\nabla_{\boldsymbol{w}^s} l_{su}(y, \boldsymbol{w}^s, \boldsymbol{\alpha}^{s\star})\right],$$
$$\nabla_{\boldsymbol{D}^s} f = \mathrm{E}\left[(\boldsymbol{x}^s - \boldsymbol{D}^s\boldsymbol{\alpha}^{s\star})\boldsymbol{\beta}^T_{\tilde{s}} - \boldsymbol{D}^s\boldsymbol{\beta}_{\tilde{s}}\boldsymbol{\alpha}^{s\star T}\right], \tag{21}$$

where $\tilde{s} = \{s, s + S, \ldots, s + (d - 1)S\}$ and $\boldsymbol{\beta} \in \mathbb{R}^{dS}$ is defined as

$$\boldsymbol{\beta}_{\Upsilon^c} = \boldsymbol{0}, \boldsymbol{\beta}_\Upsilon = (\hat{\boldsymbol{D}}^T\hat{\boldsymbol{D}} + \lambda_1\boldsymbol{\Delta} + \lambda_2\boldsymbol{I})^{-1}\boldsymbol{g}, \tag{22}$$

in which $\boldsymbol{g} = \mathrm{vec}(\nabla_{\boldsymbol{A}^\star_\Lambda}{}^T \sum_{s=1}^S l_{su}(y, \boldsymbol{w}^s, \boldsymbol{\alpha}^{s\star}))$, $\Upsilon^c = \{1, \ldots, dS\} \setminus \Upsilon$, $\boldsymbol{\beta}_\Upsilon \in \mathbb{R}^{|\Upsilon|}$ is formed of those rows of $\boldsymbol{\beta}$ indexed by Υ, and $\mathrm{vec}(.)$ is the vectorization operator.

The proof of this proposition is given in the Appendix. A stochastic gradient descent algorithm to find the optimal dictionaries $\{\boldsymbol{D}^{s\star}\}$ and classifiers $\{\boldsymbol{w}^{s\star}\}$ is described in Algorithm 1. The stochastic gradient descent algorithm is guaranteed to converge under a few assumptions that are mildly stricter than those in this paper (requires three-times differentiability) [53]. To further improve the convergence of the proposed stochastic gradient descent algorithm, a classic mini-batch strategy is used in which a small batch of the training data are sampled in each batch, instead of 1 sample, and the parameters are updated using the averaged updates of the batch. This has additional advantage in which $\hat{\boldsymbol{D}}^T\hat{\boldsymbol{D}}$ and the corresponding factorization of the ADMM for solving the sparse coding problem can be computed once for the whole batch. For the special case when $S = 1$, the proposed algorithm reduces to the single-modal task-driven dictionary learning algorithm in [29]. Selecting λ_2 in Eq. (7) to be strictly positive guarantees the linear equations of (22) to have a unique solution. In other words, it is easy to show that the matrix $(\hat{\boldsymbol{D}}^T\hat{\boldsymbol{D}} + \lambda_1\boldsymbol{\Delta} + \lambda_2\boldsymbol{I})$ is positive definite given $\lambda_1 \geq 0, \lambda_2 > 0$. However, in practice it is observed that the solution of the joint sparse representation problem is numerically stable since $\hat{\boldsymbol{D}}$ becomes full-column rank when sparsity is sought with a sufficiently large λ_1, and λ_2 can be set to zero. It should be noted that the assumption of $\hat{\boldsymbol{D}}$ being a full column rank matrix is a common assumption in sparse linear regression [26]. As in any non-convex optimization algorithm, if the algorithm is not initialized properly, it may yield poor performance. Similar to [29], the dictionaries $\{\boldsymbol{D}^s\}$ are initialized by the solution of the unsupervised multimodal dictionary learning algorithm. Upon assignment of the initial dictionaries, parameters $\{\boldsymbol{w}^s\}$ of the classifiers are set by solving (11) only with respect to $\{\boldsymbol{w}^s\}$ which is a convex optimization problem.

D. Extension

We now present an extension of the proposed algorithm with a more flexible structure on the sparse codes. Joint sparse representation relies on the fact that all the modalities share the same sparsity pattern in which, if a multimodal training sample is selected to reconstruct the input, then all the modalities within that training sample are active. However, this group sparsity constraint, imposed by the ℓ_{12} norm, may be too stringent for some applications [45], [54], for example in the scenarios where the modalities have different noise levels or when the heterogeneity of the modalities imposes different sparsity levels for the reconstruction task. A natural relaxation to the joint sparsity prior is to let the multimodal inputs not share the full active set which can be achieved by replacing the ℓ_{12} norm with a combination of the ℓ_{12} and ℓ_{11} norms ($\ell_{12} - \ell_{11}$ norm). Following the same formulation as in Section III-B, let $\boldsymbol{A}^\star(\{\boldsymbol{x}^s, \boldsymbol{D}^s\})$ in Eq. (11) be the minimizer of the

Algorithm 1 Stochastic gradient descent algorithm for multi-modal task-driven dictionary learning.

Input: Regularization parameters $\lambda_1, \lambda_2, \nu$, learning rate parameters ρ, t_0, number of iterations T, initial dictionaries $\{\boldsymbol{D}^s \in \mathcal{D}^s\}_{s \in \mathcal{S}}$, initial model parameters $\{\boldsymbol{w}^s \in \mathcal{W}^s\}_{s \in \mathcal{S}}$.

Output: Learned $\{\boldsymbol{D}^s, \boldsymbol{w}^s\}$

1: **for** $t = 1, \ldots, T$ **do**
2:     Draw a random sample $(\boldsymbol{x}_t^1, \ldots, \boldsymbol{x}_t^S, y_t)$ from the training data.
3:     Find solution $\boldsymbol{A}^\star = [\boldsymbol{\alpha}^{\star 1} \ldots \boldsymbol{\alpha}^{\star S}] \in \mathbb{R}^{d \times S}$ of the joint sparse coding problem

$$\underset{\boldsymbol{A} = [\boldsymbol{\alpha}^1 \ldots \boldsymbol{\alpha}^S]}{\operatorname{argmin}} \frac{1}{2} \sum_{s=1}^{S} \|\boldsymbol{x}_t^s - \boldsymbol{D}^s \boldsymbol{\alpha}^s\|_{\ell_2}^2 + \lambda_1 \|\boldsymbol{A}\|_{\ell_{12}} + \frac{\lambda_2}{2} \|\boldsymbol{A}\|_F^2.$$

4:     Compute set of active rows Λ of $\boldsymbol{A}^\star$ using (18).
5:     Compute $\hat{\boldsymbol{D}} \in \mathbb{R}^{n \times |\Upsilon|}$ using (19).
6:     Compute $\boldsymbol{\Delta} \in \mathbb{R}^{|\Upsilon| \times |\Upsilon|}$ using (20).
7:     Compute $\boldsymbol{\beta} \in \mathbb{R}^{dS}$ as:

$$\boldsymbol{\beta}_{\Upsilon^c} = \boldsymbol{0}, \boldsymbol{\beta}_\Upsilon = (\hat{\boldsymbol{D}}^T \hat{\boldsymbol{D}} + \lambda_1 \boldsymbol{\Delta} + \lambda_2 \boldsymbol{I})^{-1} \boldsymbol{g},$$

    where $\Upsilon = \cup_{j \in \Lambda} \{j, j + d, \ldots, j + (S-1)d\}$ and $\boldsymbol{g} = \operatorname{vec}(\nabla_{\boldsymbol{A}_{\Lambda \to}^\star}^T \sum_{s=1}^{S} l_{su}(y_t, \boldsymbol{w}^s, \boldsymbol{\alpha}^{s\star}))$.
8:     Choose the learning rate $\rho_t \leftarrow \min(\rho, \rho \frac{t_0}{t})$.
9:     Update the parameters by a projected gradient step:

$$\boldsymbol{w}^s \leftarrow \Pi_{\mathcal{W}^s} [\boldsymbol{w}^s - \rho_t (\nabla_{\boldsymbol{w}^s} l_{su}(y_t, \boldsymbol{w}^s, \boldsymbol{\alpha}^{s\star}) + \nu \boldsymbol{w}^s)],$$

$$\boldsymbol{D}^s \leftarrow \Pi_{\mathcal{D}^s} \left[\boldsymbol{D}^s - \rho_t \left((\boldsymbol{x}_t^s - \boldsymbol{D}^s \boldsymbol{\alpha}^{s\star}) \boldsymbol{\beta}_{\tilde{s}}^T - \boldsymbol{D}^s \boldsymbol{\beta}_{\tilde{s}} \boldsymbol{\alpha}^{s\star T} \right) \right],$$

    $\forall s \in \mathcal{S}$, where $\tilde{s} = \{s, s + S, \ldots, s + (d-1)S\}$.
10: **end for**

---

following optimization problem:

$$\min_{\boldsymbol{A}} \frac{1}{2} \sum_{s=1}^{S} \|\boldsymbol{x}^s - \boldsymbol{D}^s \boldsymbol{\alpha}^s\|_{\ell_2}^2 \qquad (23)$$
$$+ \lambda_1 \|\boldsymbol{A}\|_{\ell_{12}} + \lambda_1' \|\boldsymbol{A}\|_{\ell_{11}} + \frac{\lambda_2}{2} \|\boldsymbol{A}\|_F^2,$$

where λ_1' is the regularization parameter for the added ℓ_{11} norm and other terms are the same as those in Eq. (7). The selection of λ_1 and λ_1' influences the sparsity pattern of $\boldsymbol{A}^\star$. Intuitively, as λ_1/λ_1' increases, the group constraint becomes dominant and more collaboration is enforced among the modalities. On the other hand, small values of λ_1/λ_1' encourage independent reconstructions across modalities. In the extreme case of λ_1 being set to zero, the above optimization problem is separable across the modalities. The above formulation brings added flexibility with the cost of one additional design parameter which is obtained in this paper using cross-validation.

Here we present how the Algorithm 1 should be modified to solve the supervised multimodal dictionary learning problem under the mixed $\ell_{12} - \ell_{11}$ constraint. The proof for obtaining the algorithm is similar to the one for the ℓ_{12} norm and is briefly discussed in the appendix. In Algorithm 1, let $\boldsymbol{A}^\star$ be the solution of the optimization problem (23) and let Λ be the set of its active rows. Let $\Psi \subseteq \{1, \ldots, S|\Lambda|\}$ be the set of indices with non-zero entries in $\operatorname{vec}(\boldsymbol{A}_{\Lambda \to}^{\star T})$; i.e. it consists of non-zero entries of the active rows of $\boldsymbol{A}^\star$. Let $\hat{\boldsymbol{D}}, \boldsymbol{\Delta}$, and $\boldsymbol{g}$ be the same as those defined in algorithm 1. Then, $\boldsymbol{\beta} \in \mathbb{R}^{dS}$ is updated as

$$\boldsymbol{\beta}_{\Upsilon^c} = \boldsymbol{0}, \boldsymbol{\beta}_\Upsilon = (\hat{\boldsymbol{D}}_\Psi^T \hat{\boldsymbol{D}}_\Psi + \lambda_1 \boldsymbol{\Delta}_{\Psi \to, \Psi} + \lambda_2 \boldsymbol{I})^{-1} \boldsymbol{g}_\Psi,$$

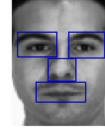
Fig. 2: Extracted modalities from a sample in AR dataset.

where Υ is the set of indices with non-zero entries in $\operatorname{vec}(\boldsymbol{A}^{\star T})$ and $\Upsilon^c = \{1, \ldots, dS\} \setminus \Upsilon$. Note that Υ is defined over the entire matrix $\boldsymbol{A}^\star$ while Ψ is defined over its active rows. The rest of the algorithm remains unchanged.

IV. RESULTS AND DISCUSSION

The performance of the proposed multimodal dictionary learning algorithms are evaluated on the AR face database [55], the CMU Multi-PIE dataset [56], the IXMAS action recognition dataset [57] and the WVU multimodal dataset [58]. For these algorithms, l_s is chosen to be the quadratic loss of Eq. (16) to handle the multiclass classification. In our experiments, it is observed that using the multiclass formulation achieves similar classification performance compared to using the logistic loss formulation of Eq. (13) in the one-vs-all setting. Regularization parameters λ_1 and ν are selected using cross-validation in the sets $\{0.01 + 0.005k | k \in \{-3, 3\}\}$ and $\{10^{-2}, \ldots, 10^{-9}\}$, respectively. It is observed that when the number of dictionary atoms is kept small compared to the number of training samples, ν can be arbitrarily set to a small value, e.g. $\nu = 10^{-8}$, for the normalized inputs. When the mixed $\ell_{12} - \ell_{11}$ norm is used, the regularization parameters λ_1 and λ_1' are selected by cross-validation in the set $\{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05\}$. The parameter λ_2 is set to zero in most of the experiments except when using the ℓ_{11} prior in Section IV-B1 where a small positive value for λ_2 was required for convergence. The learning parameter ρ_t is selected according to the heuristic proposed in [29], i.e. $\rho_t = \min(\rho, \rho \frac{t_0}{t})$ where ρ and t_0 are constants. This results in a constant learning rate during the first t_0 iterations and an annealing strategy of $1/t$ for the rest of the iterations. It is observed that choosing $t_0 = T/10$, where T is the total number of iterations over the whole training set, works well for all of our experiments. Different values of ρ are tried during the first few iterations and the one that results in minimum error on a small validation set is retained. T is set equal to be 20 in all the experiments. We observed empirically that the selection of these parameters is quite robust and small variations in their values do not affect considerably the obtained results. We also used a mini-batch size of 100 in all our experiments. It should also be noted that design parameters for the competitive algorithms are also selected using cross-validation for a fair comparison.

A. AR face recognition

The AR dataset consists of faces under different poses, illumination and expression conditions, captured in two sessions. A set of 100 users are used, each consisting of

TABLE III: Multimodal classification results obtained for the AR datasets

SVM-Maj	SVM-Sum	LR-Maj	LR-Sum	MKL	JSRC [14]	JDSRC [7]	SMDL$_{\ell_{11}}$	SMDL$_{\ell_{12}}$	SMDL$_{\ell_{12}-\ell_{11}}$
85.57	92.14	85.00	91.14	91.14	96.14	96.14	95.86	96.86	**97.14**

TABLE I: Correct classification rates obtained using the whole face modality for the AR database.

SVM	MKL [59]	LR	SRC [10]	UDL	SDL [29]
86.43	82.86	81.00	88.86	89.58	**90.57**

TABLE II: Comparison of the ℓ_{11} and ℓ_{12} priors for multimodal classification. Modalities include 1. left periocular, 2. right periocular, 3. nose, 4. mouth, and 5. face.

Modalities	$\{1,2\}$	$\{1,2,3\}$	$\{1,2,3,4\}$	$\{1,2,3,4,5\}$
UMDL$_{\ell_{11}}$	81.9	87.57	90.14	95.57
UMDL$_{\ell_{12}}$	82.6	87.86	92.00	96.29
SMDL$_{\ell_{11}}$	83.86	**89.86**	92.42	95.86
SMDL$_{\ell_{12}}$	**86.43**	**89.86**	**93.57**	**96.86**

TABLE IV: Comparison of the reconstructive-based (JSRC and JSRC-UDL) and the proposed discriminative-based (SMDL$_{\ell_{12}}$) classification algorithms obtained using the joint sparsity prior for different numbers of dictionary atoms per class on the AR dataset.

atoms/class	JSRC	JSRC-UDL	SMDL$_{\ell_{12}}$
1	46.14	71.71	**91.28**
2	69.00	78.86	**95.00**
3	79.57	83.57	**95.71**
4	88.14	91.14	**96.86**
5	91.00	94.85	**97.14**
6	94.43	96.28	**96.71**
7	**96.14**	**96.14**	96.00

seven images from the first session as training samples and seven images from the second session as test samples. A small randomly selected portion of the training set, 50 out of 700, is used as validation set for optimizing the design parameters. Fusion is taken on five modalities which are the left and right periocular, nose, mouth, and the whole face modalities, similar to the setup in [14], [45]. A test sample from the AR dataset and the extracted modalities are shown in Fig. 2. Raw pixels are first PCA-transformed and then normalized to have zero mean and unit l_2 norm. The dictionary size for the dictionary learning algorithms is chosen to be four per class, resulting in dictionaries of overall 400 atoms.

Classification using the whole face modality: The classification results using the whole face modality are shown in Table I. The results are obtained using linear support vector machine (SVM) [60], multiple kernel learning (MKL) [59], logistic regression (LR) [60], sparse representation classification (SRC) [10], and unsupervised and supervised dictionary learning algorithms (UDL and SDL) [29]. For the MKL algorithm, linear, polynomial, and RBF kernels are used. The UDL and SDL are equipped with the quadratic classifier (16). The SDL results in the best performance.

ℓ_{11} *vs* ℓ_{12} *sparse priors for multimodal classification*: A straightforward way of utilizing the single-modal dictionary learning algorithms, namely UDL and SDL, for multimodal classification is to train independent dictionaries and classifiers for each modality and then combine the individual scores for a fused decision. This way of fusion is equivalent to using the ℓ_{11} norm on $\boldsymbol{A}$, instead of ℓ_{12} norm, in Eq. (7) (or setting λ_1 to zero in Eq. (23)) which does not enforce row sparsity in the sparse coefficients. We denote the corresponding unsupervised and supervised multimodal dictionary learning algorithms using only the ℓ_{11} norm as UMDL$_{\ell_{11}}$ and SMDL$_{\ell_{11}}$, respectively. Similarly, the proposed unsupervised and supervised multimodal dictionary learning algorithms using the ℓ_{12} norm are denoted as UMDL$_{\ell_{12}}$ and SMDL$_{\ell_{12}}$. Table II compares the performance of the multimodal dictionary learning algorithms under the two priors. As shown, the proposed algorithms with ℓ_{12} prior, which enforces collaborations among the modalities, have better fusion performances than those with ℓ_{11} prior. In particular, SMDL$_{\ell_{12}}$ has significantly better performance than the SMDL$_{\ell_{11}}$ for fusion of the first and second (left and right periocular) modalities. This agrees with the intuition that these modalities are highly correlated and learning the multimodal dictionaries jointly indeed improves the recognition performance.

Comparison with other fusion methods: The performances of the proposed fusion algorithms under different sparsity priors are compared with those of the several state-of-the-art decision-level and feature-level fusion algorithms. In addition to ℓ_{11} and ℓ_{12} priors, we evaluate the proposed supervised multimodal dictionary learning algorithm with the mixed $\ell_{12}-\ell_{11}$ norm which is denoted as SMDL$_{\ell_{12}-\ell_{11}}$. One way to achieve decision-level fusion is to train independent classifiers for each modality and aggregate the outputs by either adding the corresponding scores of each modality to come up with the fused decision, or using the majority voting among the independent decisions obtained from different modalities. These approaches are abbreviated with *Sum* and *Maj*, respectively, and are used with SVM and LR classifiers for decision-level fusion. The proposed methods are also compared with feature-level fusion methods including the joint sparse representation classifier (JSRC) [14], joint dynamic sparse representation classifier (JDSRC) [7], and MKL. For the JSRC and JDSRC, the dictionary consists of all the training samples. Table III compares the performance of our proposed algorithms with the other fusion algorithms for the AR dataset. As expected, the multimodal fusion results in significant performance improvement compared to using only the whole face modality. Moreover, the proposed SMDL$_{\ell_{12}}$ and SMDL$_{\ell_{12}-\ell_{11}}$ achieve the superior performances.
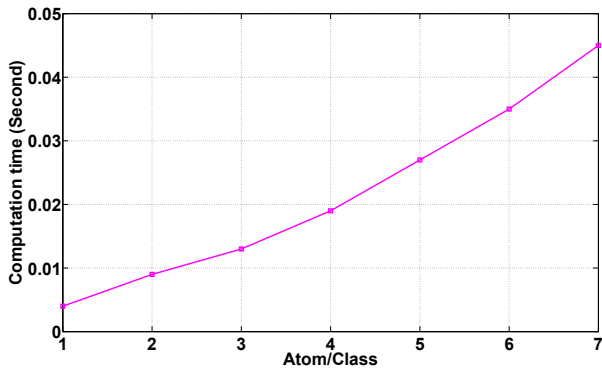
Reconstructive vs discriminative formulation with joint

Fig. 3: Computational time required to solve the optimization problem (7) for a given test sample.

Fig. 4: Configurations of the cameras and sample multi-view images from CMU Multi-Pie dataset.

TABLE V: Comparison of the supervised multimodal dictionary learning algorithms with different sparsity priors for face recognition under occlusion on the AR dataset.

$\text{SMDL}_{\ell_{12}}$	$\text{SMDL}_{\ell_{11}}$	$\text{SMDL}_{\ell_{12}-\ell_{11}}$
89.00	90.54	**91.15**

sparsity prior: Comparison of the algorithms with joint sparsity priors in Table III indicates that the proposed $\text{SMDL}_{\ell_{12}}$ algorithm equipped with dictionaries of size 400 achieves relatively better results than the JSRC that uses dictionaries of size 700. The results confirm the idea that by using the supervised formulation, compared to using the reconstruction error, one can achieve better classification performance even with more compact dictionaries. For further comparison, an experiment is performed in which the correct classification rates of the reconsturtive and discriminative formulations are compared when the their dictionary sizes are kept equal. For a given number of dictionary atoms per class d, dictionaries of JSRC are thus constructed by random selection of d train samples from different classes. This is different from the standard JSRC, utilized for the results in Table III, in which all the training samples are used to construct the dictionaries [14]. Moreover, to utilize all the available training samples for the reconstructive approach and make a more meaningful comparison, we use the unsupervised multimodal dictionary learning algorithm of Eq. (8) to train class-specific sub-dictionaries which minimizes the reconstruction error in approximating the training samples for a given class. These sub-dictionaries are then stacked to construct the final dictionaries, similar to the approach in [22]. We call this algorithm as JSRC-UDL to indicate that the dictionaries are indeed learned by the reconstructive formulation. Table IV summarizes the recognition performance of JSRC and JSRC-UDL in comparison to the proposed $\text{SMDL}_{\ell_{12}}$, which enjoys a discriminative formulation, for different number of dictionary atoms per class. As seen, $\text{SMDL}_{\ell_{12}}$ outperforms the reconstructive approaches, especially when the number of dictionary is chosen to be relatively small. This is the main advantage of $\text{SMDL}_{\ell_{12}}$ compared to the reconst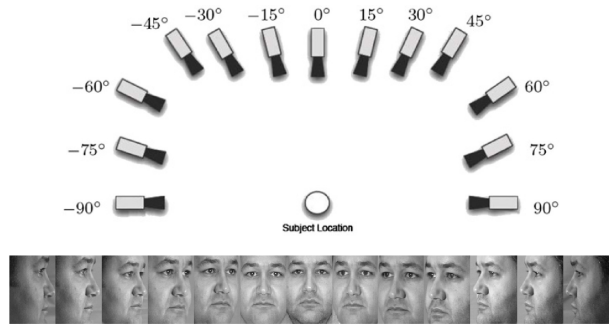ructive approaches in which more compact dictionaries can be used for the recognition task that is important for the real-time applications. It is clear that reconstructive model can only result in comparable performance when the dictionary size is chosen to be relatively large. On the other hand, the $\text{SMDL}_{\ell_{12}}$ algorithm may get over-fitted with the large number of dictionary atoms. In terms of computational expense at test time, as discussed in [14], the time required to solve the optimization problem (7) is expected to be linear in the dictionary size using the efficient ADMM if the required matrix factorization is cashed beforehand. Typical computational time to solve (7) for a given multimodal test sample is shown in Fig. 3 for different dictionary sizes. As expected, it increases linearly as the size of the dictionary increases. This illustrates the advantage of the $\text{SMDL}_{\ell_{12}}$ algorithm that results in the state-of-the-art performance with more compact dictionaries.

Classification in presence of disguise: The AR dataset also contains 600 occluded samples per session, overall 1200 images, where the faces are disguised using sun glasses or scarf. Here we use these additional images to evaluate the robustness of the proposed algorithms. Similar to previous experiments, images from session 1 are used as training samples and images from session 2 are used as test data. Classification performance under different sparsity priors are shown in Table V and as expected, the $\text{SMDL}_{\ell_{12}-\ell_{11}}$ achieves the best performance. In presence of occlusion, some of the modalities are less coupled and the joint sparsity prior among all the modalities may be too stringent as is also reflected in the results.

B. Multi-view recognition

1) Multi-view face recognition: In this section, the performance of the proposed algorithm is evaluated for multi-view face recognition using the CMU Multi-PIE dataset [56]. The dataset consists of a large number of face images under different illuminations, viewpoints, and expressions which are recorded in four sessions over the span of several months. Subjects were imaged using 13 cameras at different view-angles of $\{0°, \pm15°, \pm30°, \pm45°, \pm60°, \pm75°, \pm90°\}$ at head height. Illustrations for the multiple camera configurations, as well as sample multi-view images are shown in Fig. 4. We use the multi-view face images for 129 subjects that are

TABLE VI: Correct classification rates obtained using individual modalities in the CMU Multi-PIE database.

View	SVM	MKL	LR	SRC	UDL	SDL
Left	47.30	**52.85**	43.65	49.85	47.80	50.45
Frontal	41.15	54.10	45.40	54.25	52.10	**56.10**
Right	47.30	51.85	42.85	**52.55**	43.10	48.50

TABLE VII: Correct classification rates (CCR) obtained using multi-view images on the CMU Multi-PIE database.

Algorithm	CCR	Algorithm	CCR
SVM-Maj	62.95	LR-Maj	69.40
SVM-Sum	69.30	LR-Sum	71.10
MKL	72.40	JSRC	73.30
JDSRC	70.20	UMDL$_{\ell_{11}}$	74.80
SMDL$_{\ell_{11}}$	77.25	UMDL$_{\ell_{12}}$	70.50
SMDL$_{\ell_{12}}$	76.10	SMDL$_{\ell_{12}-\ell_{11}}$	**81.30**

Fig. 5: Sample frames of the IXMAS dataset from 5 different views.

TABLE VIII: Correct classification rates (CCR) obtained for multi-view action recognition on the IXMAS database.

Algorithm	CCR	Algorithm	CCR
Junejo et al. [65]	79.6	Tran and Sorokin [62]	80.2
Wu et al. [66]	88.2	Wang et al. 1 [63]	87.8
Wang et al. 2 [63]	93.6	JSRC	93.6
UMDL$_{\ell_{11}}$	90.3	SMDL$_{\ell_{11}}$	93.9
UMDL$_{\ell_{12}}$	90.6	SMDL$_{\ell_{12}}$	**94.8**

present in all sessions. The face regions for all the poses are extracted manually and resized to 10×8. Similar to the protocol used in [61], images from session 1 at views $\{0°, \pm30°, \pm60°, \pm90°\}$ are used as training samples. Test images are obtained from all available view angles from session 2 to have a more realistic scenario in which not all the testing poses are available in the training set. To handle multi-view recognition using the multi-modal formulation, we divide the available views into three sets of $\{-90°, -75°, -60°, -45°\}$, $\{-30°, -15°, 0°, 15°, 30°, \}$, $\{45°, 60°, 75°, 90°\}$, each of which forms a modality. A test sample is then constructed by randomly selecting an image from each modality. Two thousand test samples are generated in this way. The dictionary size for the dictionary learning algorithms is chosen to have two atoms per class.

The classification results obtained using individual modalities are shown in Table VI. As expected, better classification performance is obtained using the frontal view. Results of the multi-view face recognition is shown in Table VII. The proposed supervised dictionary learning algorithms outperform the corresponding unsupervised methods and other fusion algorithms. The SMDL$_{\ell_{12}-\ell_{11}}$ results in the state-of-the-art performance. It is consistently observed in all the studied applications that the multimodal dictionary learning algorithm with the mixed prior results in better performance than those with individual ℓ_{12} or ℓ_{12} prior. However, it requires one additional regularizing parameter to be tuned. For the rest of the paper, the performance of the proposed dictionary learning algorithms are only reported under the individual priors.

2) Multi-view action recognition: This section presents the results of the proposed algorithm for the purpose of multi-view action recognition using the IXMAS dataset [57]. Each action is recorded simultaneously by cameras from five different viewpoints, which are considered as modalities in this experiment. A multimodal sample of the IXMAS dataset is shown in Fig. 5. The dataset contains 11 action classes where each action is repeated three times by each of the ten actors, resulting in 330 sequences per view. The dataset include actions

such as check watch, cross arms, and scratch head. Similar to the work in [57], [62], [63], leave-one-actor-out cross-validation is performed and samples from all five views are used for training and testing.

We use dense trajectories as features which are generated using the publicly available code [63] in which a 2000 word codebook is generated by a random subset of these trajectories and the k-means clustering as in [64]. Note that Wang et al. [63] used HOG, HOF, and MBH descriptors in addition to the dense trajectories. However, here only dense trajectory descriptors are used. The number of dictionary atoms for the proposed dictionary learning algorithms are chosen to be 4 atoms per class, resulting in a dictionary of 44 atoms per view. The five dictionaries for JSRC are constructed using all the training samples, thus each dictionary, corresponding to a different view, has 297 atoms.

Table VIII shows average accuracies over all classes obtained using the existing algorithms and the state of the art algorithms. The Wang et al. 1 [63] algorithm uses only the dense trajectories as feature, similar to our setup. The Wang et al. 2 [63] algorithm, however, uses HOG, HOF, MBH descriptors and the spatio-temporal pyramids in addition to the trajectory descriptor. The results show that the proposed SMDL$_{\ell_{12}}$ algorithm achieves the superior performance while the SMDL$_{\ell_{11}}$ algorithm achieves the second best performance. This indicates that sparse coefficients generated by the trained dictionaries are indeed more discriminative than the engineered features. The resulting confusion matrix of the SMDL$_{\ell_{12}}$ algorithm is shown in Fig. 6.

C. Multimodal biometric recognition

The WVU dataset consists of different biometric modalities such as fingerprint, iris, palmprint, hand geometry, and voice from subjects of different age, gender, and ethnicity. It is a challenging data set, as many of the samples are corrupted with blur, occlusion, and sensor noise. In this paper, two irises (left and right) and four fingerprint modalities are used. The evaluation is done on a subset of 202 subjects which have more than four samples in all

TABLE IX: Correct classification rates obtained using individual modalities in the WVU database.

|	Finger 1	Finger 2	Finger 3	Finger 4	Iris 1	Iris 2
SVM	56.77 ± 0.72	82.95 ± 2.15	55.83 ± 2.03	80.47 ± 0.91	60.67 ± 1.78	57.52 ± 1.95
MKL	61.81 ± 1.39	82.55 ± 1.47	63.50 ± 1.75	81.85 ± 0.74	56.31 ± 2.20	54.49 ± 0.79
LR	55.64 ± 1.89	81.10 ± 1.85	55.21 ± 2.21	78.82 ± 0.66	55.25 ± 1.48	56.86 ± 1.70
SRC	$\mathbf{67.66 \pm 1.86}$	88.68 ± 1.59	$\mathbf{69.29 \pm 0.77}$	$\mathbf{88.68 \pm 1.03}$	65.43 ± 1.24	$\mathbf{67.78 \pm 1.76}$
UDL	64.68 ± 2.11	87.35 ± 2.23	67.35 ± 1.22	86.40 ± 0.70	64.36 ± 1.37	65.23 ± 2.02
SDL	66.29 ± 1.81	$\mathbf{88.84 \pm 2.31}$	68.61 ± 1.30	87.50 ± 0.82	$\mathbf{66.05 \pm 0.75}$	67.31 ± 1.38
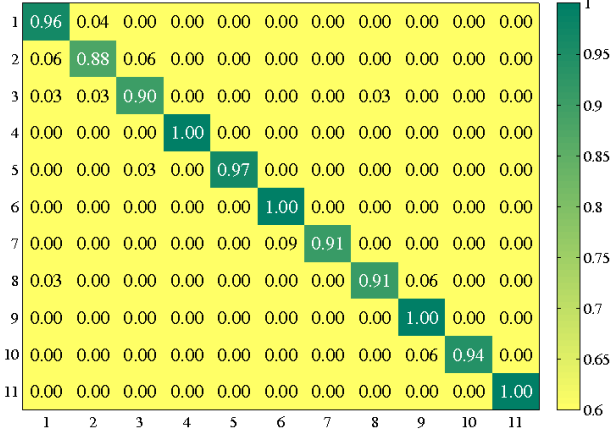
Fig. 6: The confusion matrix obtained by the $\text{SMDL}_{\ell_{12}}$ algorithm on the IXMAS dataset. The actions are 1: check watch, 2: cross arms, 3: scratch head, 4: sit down, 5: get up, 6: turn around, 7: walk, 8: wave, 9: punch, 10: kick and 11: pick up.

TABLE X: Multimodal classification results obtained for the WVU dataset.

Algorithm	4 Fingerprints	2 Irises	All modalities
SVM-Maj	90.14 ± 0.70	65.30 ± 1.92	95.24 ± 0.92
SVM-Sum	93.56 ± 1.26	74.03 ± 1.89	97.09 ± 0.83
LR-Maj	89.23 ± 1.63	63.73 ± 1.29	94.18 ± 1.13
LR-Sum	93.60 ± 0.96	71.43 ± 1.91	98.51 ± 0.18
MKL	93.28 ± 1.52	67.23 ± 0.70	94.46 ± 0.87
JSRC	$\mathbf{97.64 \pm 0.44}$	82.94 ± 0.78	98.89 ± 0.30
JDSRC	97.17 ± 0.26	79.61 ± 0.70	97.80 ± 0.51
$\text{UMDL}_{\ell_{11}}$	97.09 ± 0.56	80.90 ± 0.61	98.62 ± 0.46
$\text{SMDL}_{\ell_{11}}$	97.41 ± 0.71	82.83 ± 0.87	98.66 ± 0.43
$\text{UMDL}_{\ell_{12}}$	96.78 ± 0.57	81.53 ± 2.18	98.78 ± 0.43
$\text{SMDL}_{\ell_{12}}$	97.56 ± 0.41	$\mathbf{83.77 \pm 0.89}$	$\mathbf{99.10 \pm 0.30}$

modalities. Samples from different modalities are shown in Fig. 1. The training set is formed by randomly selecting four samples from each subject, overall 808 samples. The remaining 509 samples are used for testing. The features used here are those described in [14] which are further PCA-transformed. The dimension of the input data after preprocessing are 178 and 550 for the fingerprint and iris modalities, respectively. All inputs are normalized to have zero mean and unit l_2 norm. The number of dictionary atoms for the dictionary learning algorithms are chosen to be 2 per class, resulting in dictionaries of overall 404 atoms. The dictionaries for JSRC and JDSRC are constructed using all the training samples.

The classification results obtained using individual modalities on 5 different splits of the data into training and test samples are shown in Table IX. As shown, finger 2 is the strongest modality for the recognition task. The SRC and SDL algorithms achieve the best results. It should be noted that dictionary size of SRC is twice of that in SDL.

For multimodal classification, we consider fusion of fingerprints, fusion of Irises, and fusion of all the modalities. Table X summarizes the correct classification rates of several fusion algorithms using 4 fingerprints, 2 Irises, and all the modalities, obtained on 5 different training and test splits. Fig. 7 shows the corresponding cumulative matched score curves (CMC) for the competitive methods. CMC is a performance measure, similar to ROC, which is origi-

nally proposed for biometric recognition systems [67]. As seen, the $\text{SMDL}_{\ell_{12}}$ algorithm outperforms the competitive algorithms and achieves the state-of-the-art performance using the Irises and all modalities with the rank one recognition rate of 83.77% and 99.10%, respectively. Using the fingerprints, the performance of the $\text{SMDL}_{\ell_{12}}$ is close to the best performing algorithm, which is JSRC. The results suggest that using joint sparsity prior indeed improves the multimodal classification performance by extracting the coupled information among the modalities.

Comparison of the algorithms with the joint sparsity priors indicates that the proposed $\text{SMDL}_{\ell_{12}}$ algorithm equipped with dictionaries of size 404 achieves comparable, and mostly better, results than the JSRC that uses dictionary of size 808. Similar to the experiment in Section IV-A, we compared the reconstructive and discriminating algorithms that are based on the joint sparsity prior when the number of dictionary atoms per class is kept equal. Fig. 8 summarizes the results of the different fusion scenarios. As seen, $\text{SMDL}_{\ell_{12}}$ significantly outperforms JSRC and JSRC-UDL when the number of dictionary atoms per class is chosen to be 1 or 2. The results are consistent with that of Table IV for the AR dataset indicating that the proposed supervised formulation equipped with more compact dictionaries achieves superior performance than that of the reconstructive formulation for the studied biometric recognition applications.

V. CONCLUSIONS AND FUTURE WORKS

The problem of multimodal classification using sparsity models was studied and a task-driven formulation was proposed to jointly find the optimal dictionaries and classifiers under the joint sparsity prior. It was shown that the
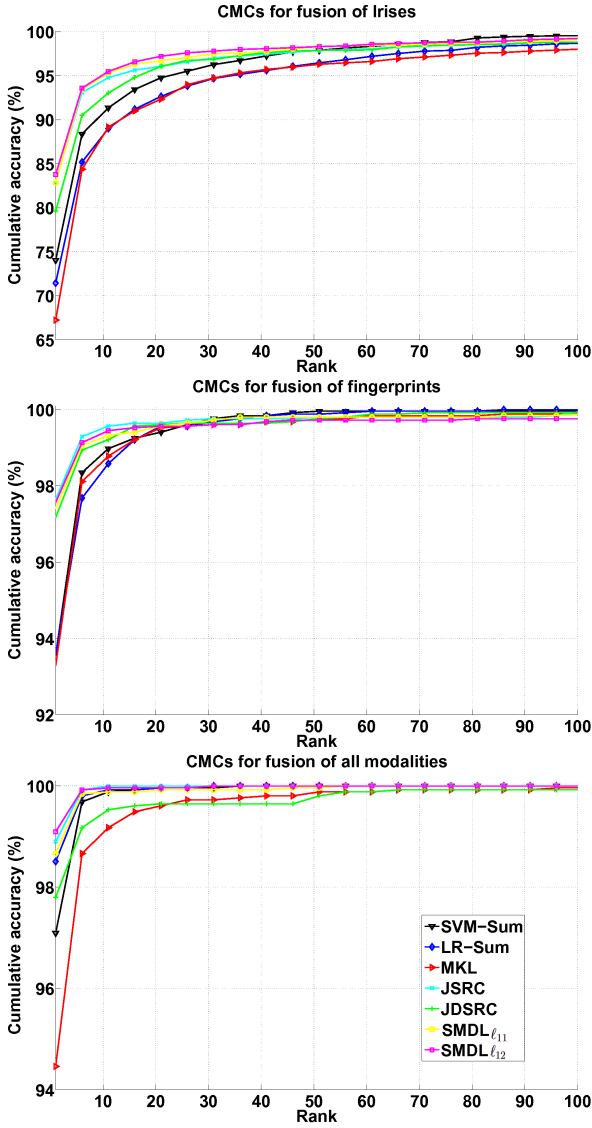
Fig. 7: CMC plots obtained by fusing the Irises (top), fingerprints (middle), and all modalities (below) on the WVU dataset.

resulting bi-level optimization problem is smooth and an stochastic gradient descent algorithm was proposed to solve the corresponding optimization problem. The algorithm was then extended for a more general scenario where the sparsity prior was the combination of the joint and independent sparsity constraints. The simulation results on the studied image classification applications suggest that while the unsupervised dictionaries can be used for feature learning, the sparse coefficients generated by the proposed multimodal task-driven dictionary learning algorithms are usually more discriminative and therefore can result in improved multimodal classification performance. It was also shown that, compared to the sparse-representation classification algorithms (JSRC, JDSRC, and JSRC-UDL), the proposed algorithms can achieve significantly better performance when compact dictionaries are utilized.

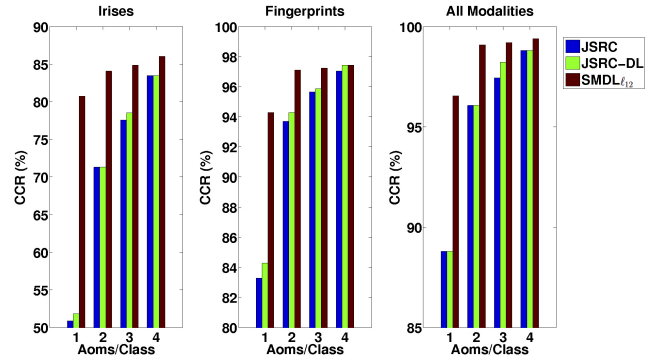In the proposed dictionary learning framework which

Fig. 8: Comparison of the reconstructive-based (JSRC and JSRC-UDL) and the proposed discriminative-based (SMDL$_{\ell_{12}}$) classification algorithms obtained using the joint sparsity prior for different numbers of dictionary atoms per class on the WVU dataset.

utilizes the stochastic gradient algorithm, the learning rate should be carefully chosen for convergence of the algorithm. In out experiments, a heuristic was used to control the learning rate. Topics of future research include developing of better optimization tools for fast convergence guarantee in this non-convex setting. Moreover, developing task-driven dictionary learning algorithms under other proposed structured sparsity priors for multimodal fusion such as the tree-structured sparsity prior [45], [68] is another future research topic. Future research will also include adapting of the proposed algorithms for other multimodal tasks such as multimodal retrieval, multimodal action recognition using Kinect data, and image super-resolution.

APPENDIX

The proof of Proposition 3.1 is presented using the following two results.

Lemma A.1 (Optimality condition): The matrix $\boldsymbol{A}^\star = \left[\boldsymbol{\alpha}^{1\star} \ldots \boldsymbol{\alpha}^{S\star}\right] = \left[\boldsymbol{a}_{1\to}^{\star\,T} \ldots \boldsymbol{a}_{d\to}^{\star\,T}\right]^T \in \mathbb{R}^{d\times S}$ is a minimizer of (7) if and only if , $\forall j \in \{1,\ldots,d\}$,

$$\begin{cases} \left[\boldsymbol{d}_j^{1\,T}\left(\boldsymbol{x}^1 - \boldsymbol{D}^1\boldsymbol{\alpha}^{1\star}\right) \ldots \boldsymbol{d}_j^{S\,T}\left(\boldsymbol{x}^S - \boldsymbol{D}^S\boldsymbol{\alpha}^{S\star}\right)\right] \\ \quad -\lambda_2 \boldsymbol{a}_{j\to}^\star = \lambda_1 \dfrac{\boldsymbol{a}_{j\to}^\star}{\|\boldsymbol{a}_{j\to}^\star\|_{\ell_2}}, \text{ if } \|\boldsymbol{a}_{j\to}^\star\|_{\ell_2} \neq 0, \\ \left\|\left[\boldsymbol{d}_j^{1\,T}\left(\boldsymbol{x}^1 - \boldsymbol{D}^1\boldsymbol{\alpha}^{1\star}\right) \ldots \boldsymbol{d}_j^{S\,T}\left(\boldsymbol{x}^S - \boldsymbol{D}^S\boldsymbol{\alpha}^{S\star}\right)\right]\right. \\ \quad \left. -\lambda_2\boldsymbol{a}_{j\to}^\star\right\|_{\ell_2} \leq \lambda_1, \text{otherwise.} \end{cases}$$
$$(24)$$

Proof: . The proof follows directly from the subgradient optimality condition of (7), i.e.

$$\boldsymbol{0} \in \{ \left[\boldsymbol{D}^{1\,T}\left(\boldsymbol{D}^1\boldsymbol{\alpha}^{1\star} - \boldsymbol{x}^1\right) \ldots \boldsymbol{D}^{S\,T}\left(\boldsymbol{D}^S\boldsymbol{\alpha}^{S\star} - \boldsymbol{x}^S\right)\right] \\ + \lambda_2\boldsymbol{A}^\star + \lambda_1\boldsymbol{P} : \boldsymbol{P} \in \partial\|\boldsymbol{A}^\star\|_{\ell_{12}}\},$$

where $\partial\|\boldsymbol{A}^\star\|_{\ell_{12}}$ denotes the subgradient of the ℓ_{12} norm evaluated at $\boldsymbol{A}^\star$. As shown in [69], the subgradient is characterized, for all $j \in \{1,\ldots,d\}$, as $\boldsymbol{p}_{j\to} = \dfrac{\boldsymbol{a}_{j\to}}{\|\boldsymbol{a}_{j\to}\|_{\ell_2}}$ if $\|\boldsymbol{a}_{j\to}\|_{\ell_2} > 0$, and $\|\boldsymbol{p}_{j\to}\|_{\ell_2} \leq 1$ otherwise. ∎

Before proceeding to the next proposition, we need to define the term *transition point*. For a given $\{\boldsymbol{x}^s\}$, let Λ_λ be the active set of the solution $\boldsymbol{A}^\star$ of (7) when $\lambda_1 = \lambda$. Then λ is defined to be a transition point of $\{\boldsymbol{x}^s\}$ if $\Lambda_{\lambda+\epsilon} \neq \Lambda_{\lambda-\epsilon}, \forall \epsilon > 0$.

Proposition A.1 (Regularity of $\boldsymbol{A}^\star$): Let $\lambda_2 > 0$ and assumption (*A*) be hold. Then,

Part 1. $\boldsymbol{A}^\star(\{\boldsymbol{x}^s, \boldsymbol{D}^s\})$ is a continuous function of $\{\boldsymbol{x}^s\}$ and $\{\boldsymbol{D}^s\}$.

Part 2. If λ_1 is not a transition point of $\{\boldsymbol{x}^s\}$, then the active set Λ of $\boldsymbol{A}^\star(\{\boldsymbol{x}^s, \boldsymbol{D}^s\})$ is locally constant with respect to both $\{\boldsymbol{x}^s\}$ and $\{\boldsymbol{D}^s\}$. Moreover, $\boldsymbol{A}^\star(\{\boldsymbol{x}^s, \boldsymbol{D}^s\})$ is locally differentiable with respect to $\{\boldsymbol{D}^s\}$.

Part 3. $\forall \lambda_1 > 0, \exists$ a set $\mathcal{N}_{\lambda_1}$ of measure zero in which $\forall \{\boldsymbol{x}^s\} \in \{\mathbb{R}^{n_s}\}\backslash\mathcal{N}_{\lambda_1}$, λ_1 is not any of the transition points of $\{\boldsymbol{x}^s\}$.

Proof: . Part 1. In the special case of $S = 1$, which is equivalent to an elastic net problem, this has already been shown [19], [70]. Our proof follows similar steps. Assumption (*A*) guarantees that $\boldsymbol{A}^\star$ is bounded. Therefore, we can restrict the optimization problem (7) to a compact subset of $\mathbb{R}^{d\times S}$. Since $\boldsymbol{A}^\star$ is unique (imposed by $\lambda_2 > 0$) and the cost function of (7) is continuous in $\boldsymbol{A}$ and each element of the set $\{\boldsymbol{x}^s, \boldsymbol{D}^s\}$ is defined over a compact set, $\boldsymbol{A}^\star(\{\boldsymbol{x}^s, \boldsymbol{D}^s\})$ is a continuous function of $\{\boldsymbol{x}^s\}$ and $\{\boldsymbol{D}^s\}$.

Part 2 and Part 3. These statements are proved here by converting the optimization problem (7) into an equivalent group lasso problem [71] and using some recent results on it. Let the matrix $\boldsymbol{D}'_j = \text{blkdiag}(\boldsymbol{d}^1_j, \ldots, \boldsymbol{d}^S_j) \in \mathbb{R}^{n\times S}, \forall j \in \{1, \ldots, d\}$, be the block-diagnoal collection of the jth atoms of the dictionaries. Also let $\boldsymbol{D}' = [\boldsymbol{D}'_1 \ldots \boldsymbol{D}'_d] \in \mathbb{R}^{n\times Sd}$, $\boldsymbol{x}' = \begin{bmatrix} \boldsymbol{x}^{1T} \ldots \boldsymbol{x}^{ST} \end{bmatrix}^T \in \mathbb{R}^n$, and $\boldsymbol{a}' = [\boldsymbol{a}_{1\rightarrow} \ldots \boldsymbol{a}_{d\rightarrow}]^T \in \mathbb{R}^{Sd}$. Then (7) can be rewritten as

$$\min_{\boldsymbol{A}} \frac{1}{2}\|\boldsymbol{x}' - \boldsymbol{D}'\boldsymbol{a}'\|^2_{\ell_2} + \lambda_1 \sum_{j=1}^d \|\boldsymbol{a}_{j\rightarrow}\|_{\ell_2} + \frac{\lambda_2}{2}\|\boldsymbol{A}\|^2_F. \quad (25)$$

This can be further converted into the standard group lasso:

$$\min_{\boldsymbol{A}} \frac{1}{2}\|\boldsymbol{x}'' - \boldsymbol{D}''\boldsymbol{a}'\|^2_{\ell_2} + \lambda_1 \sum_{j=1}^d \|\boldsymbol{a}_{j\rightarrow}\|_{\ell_2}, \quad (26)$$

where $\boldsymbol{x}'' = \begin{bmatrix} \boldsymbol{x}'^T \boldsymbol{0}^T \end{bmatrix}^T \in \mathbb{R}^{n+Sd}$ and $\boldsymbol{D}'' = \begin{bmatrix} \boldsymbol{D}'^T \sqrt{\lambda_2}\boldsymbol{I} \end{bmatrix}^T \in \mathbb{R}^{(n+Sd)\times Sd}$. It is clear that the matrix $\boldsymbol{D}''$ is full column rank. The rest of the proof follows directly from the results in [72]. ∎

Proof of Proposition 3.1: The above proposition implies that $\boldsymbol{A}^\star$ is differentiable almost everywhere. We know prove the proposition 3.1. It is easy to show that f is differentiable with respect to $\boldsymbol{w}^s$ due to the assumption (*A*) and the fact that l_{su} is twice differentiable. f is also differentiable with respect to $\boldsymbol{D}^s$ given assumption (*A*), twice differentiability of l_{su}, and the fact that $\boldsymbol{A}^\star$ is differentiable everywhere except on a set of measure zero (Prop A.1). We obtain the derivative of f with respect to $\boldsymbol{D}^s$ using the chain rule. The steps are similar to those taken for ℓ_1-related optimization in [36], though a bit more involved. Since the active set is locally constant, using the optimality condition (24), we can implicitly differentiate $\boldsymbol{A}^\star(\{\boldsymbol{x}^s, \boldsymbol{D}^s\})$ with respect to $\boldsymbol{D}^s$. For the non-active rows of $\boldsymbol{A}^\star$, the differential is zero. On the active set Λ, (24) can be rewritten as

$$\begin{bmatrix} \boldsymbol{D}^{1T}_\Lambda \left(\boldsymbol{x}^1 - \boldsymbol{D}^1\boldsymbol{\alpha}^{1\star}\right) \ldots \boldsymbol{D}^{ST}_\Lambda \left(\boldsymbol{x}^S - \boldsymbol{D}^S\boldsymbol{\alpha}^{S\star}\right) \end{bmatrix}$$
$$- \lambda_2\boldsymbol{A}^\star_{\Lambda\rightarrow} = \lambda_1 \begin{bmatrix} \dfrac{\boldsymbol{a}^{\star}_{1\rightarrow}{}^T}{\|\boldsymbol{a}^{\star}_{1\rightarrow}\|_{\ell_2}} \ldots \dfrac{\boldsymbol{a}^{\star}_{N\rightarrow}{}^T}{\|\boldsymbol{a}^{\star}_{N\rightarrow}\|_{\ell_2}} \end{bmatrix}^T, \quad (27)$$

where N is the cardinality of Λ and $\boldsymbol{D}_\Lambda$ and $\boldsymbol{A}^\star_{\Lambda\rightarrow}$ are the matrices consisting of active columns of $\boldsymbol{D}$ and active rows of $\boldsymbol{A}^\star$, respectively. For the rest of the proof, we only work on the active set and the symbols Λ and $\star$ are dropped for the ease of notation. Taking the partial derivative from both sides of (27) with respect to d^s_{ij}, the element in the ith-row and j^{th}-column of $\boldsymbol{D}^s$, and taking its transpose we have:

$$\lambda_2 \frac{\partial \boldsymbol{A}^T}{\partial d^s_{ij}} + \begin{bmatrix} \boldsymbol{0} \\ (\boldsymbol{D}^s\boldsymbol{\alpha}^s - \boldsymbol{x}^s)^T \boldsymbol{E}^s_{ij} + \boldsymbol{\alpha}^{sT} \boldsymbol{E}^s_{ij}{}^T \boldsymbol{D}^s \\ \boldsymbol{0} \end{bmatrix} +$$
$$\begin{bmatrix} \dfrac{\partial \boldsymbol{\alpha}^{1T}}{\partial d^s_{ij}}\boldsymbol{D}^{1T}\boldsymbol{D}^1 \\ \vdots \\ \dfrac{\partial \boldsymbol{\alpha}^{ST}}{\partial d^s_{ij}}\boldsymbol{D}^{ST}\boldsymbol{D}^S \end{bmatrix} = -\lambda_1 \begin{bmatrix} \boldsymbol{\Delta}_1 \dfrac{\partial \boldsymbol{a}^T_{1\rightarrow}}{\partial d^s_{ij}} \ldots \boldsymbol{\Delta}_N \dfrac{\partial \boldsymbol{a}^T_{N\rightarrow}}{\partial d^s_{ij}} \end{bmatrix},$$

where $\boldsymbol{E}^s_{ij} \in \mathbb{R}^{n_s\times N}$ is a matrix with zero elements except the element in the ith row and j^{th} column which is one and

$$\boldsymbol{\Delta}_k = \frac{1}{\|\boldsymbol{a}_{k\rightarrow}\|_{\ell_2}} \left(\boldsymbol{I} - \frac{1}{\|\boldsymbol{a}_{k\rightarrow}\|^2_{\ell_2}}\boldsymbol{a}^T_{k\rightarrow}\boldsymbol{a}_{k\rightarrow}\right) \in \mathbb{R}^{S\times S},$$

$\forall k \in \{1, \ldots, N\}$. It is easy to check that $\boldsymbol{\Delta}_k \geq \boldsymbol{0}$. Vectorizing the both sides and factorizing results in

$$\text{vec}\left(\frac{\partial \boldsymbol{A}^T}{\partial d^s_{ij}}\right) =$$
$$\boldsymbol{P} \begin{bmatrix} \boldsymbol{0} \\ (\boldsymbol{x}^s - \boldsymbol{D}^s\boldsymbol{\alpha}^s)^T \boldsymbol{e}^s_{ij_1} - \boldsymbol{\alpha}^{sT} \boldsymbol{E}^s_{ij}{}^T \boldsymbol{d}^s_1 \\ \vdots \\ (\boldsymbol{x}^s - \boldsymbol{D}^s\boldsymbol{\alpha}^s)^T \boldsymbol{e}^s_{ij_N} - \boldsymbol{\alpha}^{sT} \boldsymbol{E}^s_{ij}{}^T \boldsymbol{d}^s_N \\ \boldsymbol{0} \end{bmatrix}, \quad (28)$$

where $\boldsymbol{e}^s_{ij_k}$ is the kth column of $\boldsymbol{E}^s_{ij}$, $\boldsymbol{P} = \left(\hat{\boldsymbol{D}}^T\hat{\boldsymbol{D}} + \lambda_1\boldsymbol{\Delta} + \lambda_2\boldsymbol{I}\right)^{-1}$, and $\hat{\boldsymbol{D}}$ and $\boldsymbol{\Delta}$ are defined in Eqs. (19) and (20), respectively. Further simplifying Eq. (28) yields

$$\text{vec}\left(\frac{\partial \boldsymbol{A}^T}{\partial d^s_{ij}}\right) = \boldsymbol{P}_{\tilde{s}}\boldsymbol{E}^s_{ij}{}^T (\boldsymbol{x}^s - \boldsymbol{D}^s\boldsymbol{\alpha}^s) - \boldsymbol{P}_{\tilde{s}}\boldsymbol{d}^s_{i\rightarrow}{}^T\alpha^s_j,$$

where $\tilde{s}$ is defined in Eq. (21). Using the chain rule, we have

$$\frac{\partial f}{\partial d^s_{ij}} = \text{E}\left[\boldsymbol{g}^T \text{vec}\left(\frac{\partial \boldsymbol{A}^T}{\partial d^s_{ij}}\right)\right],$$

where $\boldsymbol{g} = \text{vec}\left(\frac{\partial \sum_{s=1}^{S} l_{su}}{\partial \boldsymbol{A}^T}\right)$. Therefore, derivative with respective to the active columns of dictionary $\boldsymbol{D}^s$ is

$$\frac{\partial f}{\partial \boldsymbol{D}^s} = \text{E} \begin{bmatrix} \boldsymbol{g}^T \boldsymbol{P}_{\tilde{s}} \left(\boldsymbol{E}_{11}^{s\,T} (\boldsymbol{x}^s - \boldsymbol{D}^s \boldsymbol{\alpha}^s) - \boldsymbol{d}_{1\rightarrow}^{s\,T} \alpha_1^s \right) \\ \vdots \\ \boldsymbol{g}^T \boldsymbol{P}_{\tilde{s}} \left(\boldsymbol{E}_{n_s 1}^{s\,T} (\boldsymbol{x}^s - \boldsymbol{D}^s \boldsymbol{\alpha}^s) - \boldsymbol{d}_{n_s\rightarrow}^{s\,T} \alpha_1^s \right) \\ \cdots \quad \boldsymbol{g}^T \boldsymbol{P}_{\tilde{s}} \left(\boldsymbol{E}_{1N}^{s\,T} (\boldsymbol{x}^s - \boldsymbol{D}^s \boldsymbol{\alpha}^s) - \boldsymbol{d}_{1\rightarrow}^{s\,T} \alpha_N^s \right) \\ \vdots \\ \cdots \quad \boldsymbol{g}^T \boldsymbol{P}_{\tilde{s}} \left(\boldsymbol{E}_{n_s N}^{s\,T} (\boldsymbol{x}^s - \boldsymbol{D}^s \boldsymbol{\alpha}^s) - \boldsymbol{d}_{n_s\rightarrow}^{s\,T} \alpha_N^s \right) \end{bmatrix}$$

$$= \text{E}\left[(\boldsymbol{x}^s - \boldsymbol{D}^s \boldsymbol{\alpha}^s) \boldsymbol{g}^T \boldsymbol{P}_{\tilde{s}} - \boldsymbol{D}^s \boldsymbol{P}_{\tilde{s}}^T \boldsymbol{g} \alpha^{s\,T} \right].$$

Setting $\boldsymbol{\beta} = \boldsymbol{P}^T \boldsymbol{g} \in \mathbb{R}^{NS}$ and noting that $\boldsymbol{\beta}_{\tilde{s}} = \boldsymbol{P}_{\tilde{s}}^T \boldsymbol{g}$ complete the proof. ∎

Derivation of the algorithm with the mixed $\ell_{12} - \ell_{11}$ prior can be obtained similarly. For each active row $j \in \Lambda$ of $\boldsymbol{A}^\star$, the solution of the optimization problem (23) with the mixed prior, let $\Pi_j \subseteq \mathcal{S}$ be the set of active modalities which have non-zeros entries. Then the optimality condition for the active row j is

$$\left[\boldsymbol{d}_j^{1\,T} \left(\boldsymbol{x}^1 - \boldsymbol{D}^1 \boldsymbol{\alpha}^{1\star} \right) \ldots \boldsymbol{d}_j^{S\,T} \left(\boldsymbol{x}^S - \boldsymbol{D}^S \boldsymbol{\alpha}^{S\star} \right) \right]_{\Pi_j}$$

$$- \lambda_2 \boldsymbol{a}_{j\rightarrow}^\star = \lambda_1 \frac{\boldsymbol{a}_{j\rightarrow,\Pi_j}^\star}{\|\boldsymbol{a}_{j\rightarrow}^\star\|_{\ell_2}} + \lambda_1' \, \text{sign}\left(\boldsymbol{a}_{j\rightarrow,\Pi_j}^\star \right).$$

Then, the algorithm for the mixed prior can be obtained by differentiating the optimality condition, following similar steps as was shown for the ℓ_{12} prior.

REFERENCES

[1] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proc. IEEE*, vol. 85, no. 1, pp. 6–23, January 1997.

[2] P. K. Varshney, "Multisensor data fusion," in *Intell. Problem Solving. Methodologies and Approaches*. Springer Berlin Heidelberg, 2000.

[3] H. Wu, M. Siegel, R. Stiefelhagen, and J. Yang, "Sensor fusion using Dempster-Shafer theory," in *Proc. 19th IEEE Instrum. and Meas. Technol. Conf. (IMTC)*, 2002, pp. 7–12.

[4] A. Ross and R. Govindarajan, "Feature level fusion using hand and face biometrics," in *SPIE proc. series*, 2005, pp. 196–204.

[5] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," *Comput. and Inf. syst.*, vol. 7, no. 1, pp. 1–10, 2000.

[6] A. Rattani, D. R. Kisku, M. Bicego, and M. Tistarelli, "Feature level fusion of face and fingerprint biometrics," in *Proc. 1st IEEE Int. Conf. Biometrics: Theory, Applicat., and Syst.*, 2007, pp. 1–6.

[7] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Multi-observation visual recognition via joint dynamic sparse representation," in *Proc. IEEE Conf. Comput. Vision (ICCV)*, 2011, pp. 595–602.

[8] A. Klausne, A. Tengg, and B. Rinner, "Vehicle classification on multi-sensor smart cameras using feature- and decision-fusion," in *Proc. 1st ACM/IEEE Int. Conf. Distributed Smart Cameras*, 2007, pp. 67–74.

[9] A. Rattani and M. Tistarelli, "Robust multi-modal and multi-unit feature level fusion of face and iris biometrics," in *Advances in Biometrics*, pp. 960–969. Springer, 2009.

[10] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[11] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 33, no. 11, pp. 2259–2272, Nov. 2011.

[12] H. Zhang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang, "Joint-structured-sparsity-based classification for multiple-measurement transient acoustic signals," *IEEE Trans. Syst., Man, Cybern.*, vol. 42, no. 6, pp. 1586–98, Dec. 2012.

[13] R. Caruana, *Multitask learning*, Springer, 1998.

[14] S. Shekhar, V. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 113–126, Jan. 2013.

[15] U. Srinivas, H. Mousavi, C. Jeon, V. Monga, A. Hattel, and B. Jayarao, "Simultaneous sparsity model for histopathological image representation and classification," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1163 – 1179, May 2014.

[16] H. S. Mousavi, V. Srinivas, U.and Monga, Y. Suo, M. Dao, and T. D. Tran, "Multi-task image classification via collaborative, hierarchical spike-and-slab priors," in *IEEE Intl. Conf. Image Processing (ICIP)*, Oct 2014, pp. 4236–4240.

[17] N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, "Robust multi-sensor classification via joint sparse representation," in *Proc. 14th Int. Conf. Information Fusion (FUSION)*, 2011.

[18] M. Yang, L. Zhang, D. Zhang, and S. Wang, "Relaxed collaborative representation for pattern classification," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2012, pp. 2224–2231.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," *Proc. 26th Annu. Int. Conf. Mach. Learning (ICML)*, pp. 689–696, 2009.

[20] J. Mairal, F. Bach, A. Zisserman, and G. Sapiro, "Supervised dictionary learning," in *Advances Neural Inform. Process. Syst. (NIPS)*, 2008, pp. 1033–1040.

[21] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[22] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Proc. IEEE Conf. Image Process. (ICIP)*, 2010, pp. 1601–1604.

[23] Y. L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2010, pp. 2559–2566.

[24] Z. Jiang, Z. Lin, and L. S. Davis, "Label consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 35, no. 11, pp. 2651–2664, Nov. 2013.

[25] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD : An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[26] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The J. of Mach. Learning Research*, vol. 11, pp. 19–60, 2010.

[27] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, 1999, vol. 5, pp. 2443–2446.

[28] M. Elad and M. Aharon, "Image denoising via saprse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[29] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[30] J. Yang, Z. Wang, Z. Lin, X. Shu, and T. Huang, "Bilevel sparse coding for coupled feature spaces," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2360–2367.

[31] S. Kong and D. Wang, "A brief summary of dictionary learning based approach for classification," *arxivId:1205.6544*, 2012.

[32] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[33] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2010, pp. 2691–2698.

[34] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2010, pp. 3501–3508.

[35] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2011, pp. 543–550.

[36] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2010, pp. 3517–3524.

[37] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Ann. Operations Research*, vol. 153, no. 1, pp. 235–256, Apr. 2007.

[38] D. M. Bradley and J. A. Bagnell, "Differentiable sparse coding," in *Advances Neural Inform. Process. Syst. (NIPS)*, 2008.

[39] J. Zheng and Z. Jiang, "Learning view-invariant sparse representations for cross-view action recognition," in *Proc. IEEE Intel. Conf. Computer Vision (ICCV)*, 2013, pp. 3176–3183.

[40] G. Monaci, P. Jost, P. Vandergheynst, B. Mailhé, S. Lesage, and R. Gribonval, "Learning multimodal dictionaries," *IEEE Trans. Image Process*, vol. 16, no. 9, pp. 2272–2283, Sep. 2007.

[41] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. 27th Conf. Artificial Intell.*, 2013, pp. 1070–1076.

[42] H. Zhang, Y. Zhang, and T. S. Huang, "Simultaneous discriminative projection and dictionary learning for sparse representation based classification," *Pattern Recognition*, vol. 46, no. 1, pp. 346–354, Jan. 2013.

[43] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, Dec. 1997.

[44] L. Bottou and O. Bousquet, "The trade-offs of large scale learning," in *Advances Neural Inform. Process. Syst. (NIPS)*, 2007, pp. 161–168.

[45] S. Bahrampour, A. Ray, N. M. Nasrabadi, and W. K. Jenkins, "Quality-based multimodal classification using tree-structured sparsity," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2014, pp. 4114–4121.

[46] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, Jul. 2005.

[47] J. A. Tropp, "Algorithms for simultaneous sparse approximation. part ii: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, Mar. 2006.

[48] A. Rakotomamonjy, "Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms," *Signal Process.*, vol. 91, no. 7, pp. 1505–1526, Jul. 2011.

[49] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, pp. 1–96, 2013.

[50] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Royal Statistical Soc. Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[51] M. Aharon and M. Elad, "Sparse and redundant modeling of image content using an image-signature-dictionary," *SIAM J.l on Imaging Sciences*, vol. 1, no. 3, pp. 228–247, 2008.

[52] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

[53] L. Bottou, "Online learning and stochastic approximations," *On-line learning in neural networks*, vol. 17, pp. 9.

[54] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-hilasso: A collaborative hierarchical sparse modeling framework," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4183–4198, Sep. 2011.

[55] A. M. Martinez and R. Benavente, "The AR face database," *CVC Technical Rep.*, vol. 24, 1998.

[56] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[57] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," in *IEEE 11th Intel. Conf. Computer Vision (ICCV)*, 2007, pp. 1–7.

[58] S. S. S. Crihalmeanu, A. Ross, and L. Hornak, "A protocol for multibiometric data acquisition, storage and dissemination," Tech. Rep., Lane Dept. of Comput. Sci. and Elect. Eng., West Virginia Univ., 2007.

[59] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "Simplemkl.," *J. of Mach. Learning Research*, vol. 9, no. 11, 2008.

[60] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[61] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Joint dynamic sparse representation for multi-view face recognition," *Pattern Recognition*, vol. 45, pp. 1290–1298, 2012.

[62] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *Computer Vision–ECCV 2008*, pp. 548–561. Springer, 2008.

[63] A. Wang, H.and Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Intl. J. Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[64] A. Gupta, A. Shafaei, J. J. Little, and R. J. Woodham, "Unlabelled 3d motion examples improve cross-view action recognition," in *Proc. British Machine Vision Conf.*, 2014.

[65] I. N. Junejo, E. Dexter, and P. Laptev, I.and Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, 2011.

[66] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 2011, pp. 489–496.

[67] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "The relation between the roc curve and the cmc," in *Proc. 4th IEEE Workshop Automat. Identification Advanced Technologies*, 2005, pp. 15–20.

[68] Julien Mairal, Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach, "Learning hierarchical and topographic dictionaries with structured sparsity," in *SPIE*, 2011.

[69] E. V. D. Berg and M. P. Friedlander, "Theoretical and empirical results for recovery from multiple measurements," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2516–2527, Apr. 2010.

[70] H. Zou, T. Hastie, and R. Tibshirani, "On the degrees of freedom of the lasso," *The Ann. of Statistics*, vol. 35, no. 5, pp. 2173–2192, 2007.

[71] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Soc.: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[72] S. Vaiter, C. Deledalle, G. Peyre, J. Fadili, and C. Dossal, "Degrees of freedom of the group lasso," *arXiv:1205.1481*, 2012.